Optimal Learning of Kernel Logistic Regression for Complex Classification Scenarios

Hongwei Wen, Annika Betken, Hanyuan Hang

University of Twente

Problem Setting

- **Standard Classification Scenario:** we observe i.i.d. data $\mathrm{D} := (X_i, Y_i)_{i=1}^{n}$ drawn from an unknown distribution P, where X_i denotes the input and Y_i represents the output. The goal is to predict the output Y for the input X.
- **Complex Classification Scenarios:** labeled samples $D_p := (X_i, Y_i)_{i=1}^{n_p}$ drawn from a distribution P, while inference is required for a different distribution Q on the same space.
- **Label shift assumption:** Two distributions P and Q share the same conditional probability but has different class probabilities, i.e.,

$$\mathrm{p(x|y) = q(x|y)\ but\ p(y) \neq q(y).}$$

- **Goals:** To estimate the class conditional probability (CCP) estimator

$$q(y|x) = \frac{w_y^* p(y|x)}{\sum_{k=1}^{K} w_k^* p(k|x)},$$

where the class probability ratio $w^* := (w_y^*)_{y \in [K]}$ between q(y) and p(y) is given by

$$w_y^* := \mathrm{q(y)/p(y), y} \in [\mathrm{K}].$$

We then induce the plug-in classifier defined as $argmax_{k\in[K]}\ \hat{q}(y|x)$.

Specific Tasks

- **Long-tailed Learning:** the target label probability $q(y)$ is known to be uniform, specifically $q(k) = 1/K$ for any $k \in [K]$. In contrast, the label probability $\mathrm{p}(y)$ may significantly deviate from a uniform distribution.
- **Domain Adaptation: Label Shift:** the label distribution q(y) is unknown. In addition to labeled samples D_p from the source distribution P, unlabeled samples $D_q^u \coloneqq (X_i)_{i=n_p+1}^{n_p+n_\mathrm{q}}$ are drawn from the marginal probability density q(x) of the target distribution Q.
- **Transfer Learning: Label Bias:** the label distribution q(y) is uniform. We assume that only the pre-trained model $\hat{p}(y|x)$ on D_p is available, while labeled pre-trained data D_p itself is not accessible. Additionally, we can observe a small number of auxiliary samples $D_s \coloneqq (X_i^s, Y_i^s)_{i=1}^{n_s}$, drawn from an unknown data distribution S, which is assumed to satisfy s(x|y) = p(x|y).

Methodology based on CCP estimation

- **CCP-based method:**

Step 1. estimate the class probability ratio $\widehat{w}$.

Step 2. calculate the CCP predictor of q(y|x) by

$$\hat{q}(y|x) = \frac{\hat{p}(y|x)\,\widehat{w}_y}{\sum_{k=1}^{K} \hat{p}(k|x)\,\widehat{w}_k}, \qquad \mathrm{y} \in [\mathrm{K}]. \qquad (2)$$

- **Long-tailed Learning:**

The probability \$p(y)\$ can be easily estimated by

$$\hat{p}(y) := \frac{1}{n_p} \sum_{i=1}^{n_p} 1\{Y_i = y\}, \qquad \mathrm{y} \in [\mathrm{K}]. \qquad (3)$$

Consequently, the class probability ratio w^* can be estimated as

$$\widehat{w}_y \coloneqq \hat{q}(y) / \hat{p}(y) = 1/(K\,\hat{p}(y)).$$

Methodology based on CCP estimation

- **Domain Adaptation: Label Shift:**

Wen et.al (2024) estimate **the class probability ratio** w^* by matching $\hat{p}(y)$ and the weighted conditional probability $\hat{p}(y|x)$

$$\hat{p}_q^w(y) := \frac{1}{n_q}\sum_{i=1}^{n_q}\frac{\hat{p}(\mathrm{y}|x_i)}{\sum_{\mathrm{y}=1}^{\mathrm{K}} w_y\, \hat{p}(\mathrm{y}|X_i)}.$$

Then we find the solution to the following minimization problem,

$$\hat{w} \coloneqq argmin_{w\in R^K} \sum_{y=1}^{K} |\hat{p}(y) \ - \hat{p}_q^w(y)|^2 . \qquad (4)$$

- **Transfer Learning: Label Bias**

Zhu et.al. (2024) proposes the estimator $\hat{p}(y)$ as the stationary distribution of a Markov chain characterized by the transition matrix $\hat{C} = (\hat{C}_{kj})_{k,j\in[K]}$, with entries given by

$$\hat{C}_{kj} := \frac{1}{n_{s,j}} \sum_{i=1}^{n_{s,j}} 1\{Y_i = y\}\ \hat{p}(k|X_i^s), \qquad (5)$$

where $n_{s,j}$ denotes the sample size of the j-th class in D_s. Given that $q(k) = 1/K$, the weight w_y^* can be estimated as $\hat{w}_y \coloneqq 1/(K\hat{p}(y))$.

Main Theoretical Results

Assumption 1 We impose the following assumptions on the distribution P.

- (i) *Holder Smoothness:* Assume that for any x, x', there exist a Holder constant $c_\alpha \geq 0$ and an $\alpha \in [0,1]$ such that $|p(k|x') - p(k|x) \leq c_\alpha \parallel x' - x \parallel^\alpha$ for all $k \in [K]$.
- (ii) *Small Value Bound:* Assume that for all $t \in [0,1]$, there exist constants $\beta \geq 0$ and $c_\beta \geq 0$ such that $P_X(p(k|x) \leq t) \leq c_\beta \, t^\beta$ for all $k \in [K]$.

Generalization Bounds

Theorem 2 (Generalization Bound).

Let Assumptions 1 holds. Moreover, let $\hat{q}(y|x)$ be the CCP-based estimator in Eq. (2). Then, let $\mathcal{R}_Q(\hat{q}(y|x))$ and $\mathcal{R}_Q^*$ be the cross-entropy risk of $\hat{q}(y|x)$ and the possibly minimal CE risk, respectively.

- Then for *long-tailed learning*, with probability at least $1 - 2/n_p$, for any $\xi > 0$, there holds

$$\mathcal{R}_Q(\hat{q}(y|x)) - \mathcal{R}_Q^* \preccurlyeq n_p^{-\frac{(1+\beta\wedge 1)\alpha}{(1+\beta\wedge 1)\alpha+d}+\xi}$$

- For *label shift domain adaptation*, with probability at least $1 - 2/n_p - 2/n_q$, for any $\xi > 0$, there holds

$$\mathcal{R}_Q(\hat{q}(y|x)) - \mathcal{R}_Q^* \preccurlyeq n_p^{-\frac{(1+\beta\wedge 1)\alpha}{(1+\beta\wedge 1)\alpha+d}+\xi} + \log n_q/n_q.$$

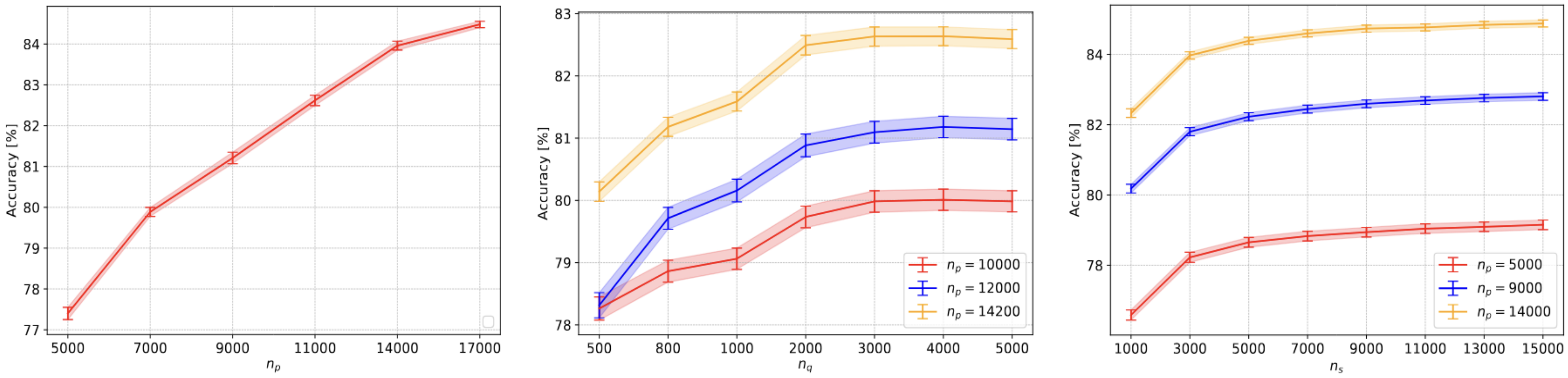
- For *label bias transfer learning*, with probability at least $1 - 2/n_p - 2/n_q$, for any $\xi > 0$, there holds

$$\mathcal{R}_Q(\hat{q}(y|x)) - \mathcal{R}_Q^* \preccurlyeq n_p^{-\frac{(1+\beta\wedge 1)\alpha}{(1+\beta\wedge 1)\alpha+d}+\xi} + \log n_s/n_s\,.$$

Empirical Verification

Dataset	Method	Long-tailed Learning	Domain Adaptation	Transfer Learning
`Dionis`	Baseline	80.71 ± 0.87	77.69 ± 1.58	80.72 ± 0.88
	CCP	**83.67 ± 1.10**	**82.73 ± 1.40**	**84.22 ± 0.99**
`Gas Sensor`	Baseline	85.57 ± 5.97	96.14 ± 0.89	85.97 ± 5.57
	CCP	**90.49 ± 4.47**	**96.52 ± 1.30**	**90.27 ± 4.78**
`Satimage`	Baseline	80.51 ± 3.70	89.80 ± 3.90	80.51 ± 3.70
	CCP	**84.56 ± 1.32**	**96.46 ± 2.42**	**84.47 ± 1.98**

For each dataset and each method, we denote the best performance with **bold**.



(a) Long-tailed learning (b) Domain adaptation (c) Transfer learning

Figure 1: The impact of sample sizes on accuracy in complex classification scenarios.